Investigating the Portability of Corpus-Derived Cue Phrases for Dialogue Act Classification

Nick Webb and **Ting Liu**
ILS Institute
University at Albany, SUNY
Albany, NY, USA
`{nwebb|tl7612}@albany.edu`

Abstract

We present recent work in the area of Cross-Domain Dialogue Act tagging. Our experiments investigate the use of a simple dialogue act classifier based on purely intra-utterance features - principally involving word n-gram cue phrases. We apply automatically extracted cues from one corpus to a new annotated data set, to determine the portability and generality of the cues we learn. We show that our automatically acquired cues are general enough to serve as a cross-domain classification mechanism.

1 Introduction

A number of researchers (Hirschberg and Litman, 1993; Grosz and Sidner, 1986) speak of *cue* or *key phrases* in utterances that can serve as useful indicators of discourse structure. We have previously investigated the use of such cue phrases to predict dialogue acts or DAs (functional tags which represent the communicative intentions behind each user utterance) (Webb et al., 2005a). We developed an approach, in common with the work of Samuel et al. (1999), where word n-grams that might serve as cue phrases are automatically detected in a corpus and we have previously reported the results of experiments evaluating this approach on the SWITCHBOARD corpus, where our results rival the best reported over that data (Stolcke et al., 2000), although our method adopts a significantly less complex algorithm.

An interesting by-product of our approach is the ranked list of cue phrases derived from the source corpus. Visual inspection of these cues reveals that, as one might expect, there is a high degree of correlation between phrases such as *"can you"* and the DA <yes/no question>, *"where is"* and *"who is"* with the DA <wh-question> and *"right"* or *"ok"* with DA <agree/accept>. These cues appear to be of a general nature, unrelated to the source domain or application. Therefore, despite being automatically acquired from one domain specific corpus, these cues should be equally applicable to new corpora, from a different domain and it is this hypothesis we test. This paper presents our work on dialogue act classification using cues automatically extracted from a corpus from one domain, and applying these cues directly as a classifier over a new corpus from a different domain.

The material is presented as follows: Previous work with dialogue act modelling is outlined in Section 2. An overview of the corpora used for the experiments we report can be seen in Section 3. A brief overview of our classification method is given in Section 4. Our experiments evaluating the cue-based dialogue act classifier tagging new, out-of-domain data are given in Section 5. Finally we end with some discussion and an outline of intended further work.

2 Related Work

Dialogue Acts (DAs) (Bunt, 1994), also known as speech acts or dialogue moves, represent the functional performance of a speaker's utterance, such as a greeting *"Hello there"*, asking a question like *"How is your mother?"* or making a request *"Can you move your foot?"*.

There are two broad categories of computational model used to interpret these acts. The first, in-

Corpus	Availability	Utterance count	Dialogue count	Word count	Distinct words	Dialogue type
SWITCHBOARD	public	223606	1155	1431725	21715	Conversational
AMITIÉS GE	restricted	30206	1000	228165	7841	Task-oriented

Figure 1: Summary data for the dialogue corpora

cluding the work of Cohen and Perrault (1979) relies on processing belief logics, centring on the impact each utterance has on the hearer - what the hearer believes the speaker intended to communicate. These models can be very accurate, but often are complex, and require significant world-knowledge to create.

The second model type is cue-based, and centres on the notion of repeated, predictive cues - subsections of language which are strong indicators of specific DAs. In this second category, much of the work is cast as a probabilistic classification task, solved by training approaches on labelled examples of dialogue acts. As an example of these probabilistic methods, Stolcke et al. (2000) apply a HMM method to the SWITCHBOARD corpus, one that exploits both the order of words *within* utterances and the order of dialogue acts *over* utterances. They use a single split of the data for their experiments, with 198k utterances for training and 4k utterances for testing, achieving a DA tagging accuracy of 71.0% on word transcripts. Another learning approach by Samuel et al. (1998) uses transformation-based learning over a number of utterance features, including utterance length, speaker turn and the dialogue act tags of adjacent utterances. They achieved an average score of 75.12% tagging accuracy over the VERBMO-BIL corpus. A significant aspect of this work is the automatic identification of word sequences that might serve as useful dialogue act cues (Samuel et al., 1999). A number of statistical criteria are applied to identify potentially useful word n-grams that are then supplied to the transformation-based learning method as 'features'.

What has been less explored is the portability or adaptability of these models to new corpora and new domains. Prasad and Walker (2002) look at applying models generated from a Human-Computer corpus to a Human-Human corpus in the same domain, that of travel planning, and score a very low 36.72% accuracy using their model. The work of Tur et al. (2006) is closer to the work reported here - they apply models derived from the

SWITCHBOARD corpus to the ICSI-MRDA corpus (Shriberg et al., 2004) using boosting, applied to a high level of representation (comprising only 5 DA categories, one of which they exclude), where they achieve 57.37% tagging accuracy. This seems to indicate that cross-domain application of models is possible, although the level of accuracy as presently reported is low.

3 Experimental Corpora

Our work as described here applies to two corpora - the DA-tagged portion of the SWITCHBOARD corpus (Jurafsky et al., 1998), and the AMITIÉS GE corpus (Hardy et al., 2002; Hardy et al., 2003), created as part of the AMITIÉS European 5th Framework program project (Hardy et al., 2005). A summary of the two corpora can be seen in Figure 1.

3.1 Switchboard

The annotated portion of the SWITCHBOARD corpus comprises 1155 annotated conversations between two human participants, where the dialogues are of an unstructured, non-directed character. Participants do not know each other, and are provided only with a set of topics which they may wish to discuss. The SWITCHBOARD corpus is annotated using an elaboration of the DAMSL tag set. In 1998 the Discourse Resource Initiative finalised a task-independent set of DAs, called DAMSL (Dialogue Act Markup in Several Layers), for use across different domains. DAMSL has been used to mark-up several dialogue corpora, such as TRAINS (Core and Allen, 1997), and the SWITCH-BOARD corpus (Jurafsky et al., 1998).

The annotation over the SWITCHBOARD corpus involves 50 major classes, together with a number of diacritic marks, which combine to generate 220 distinct labels. Jurafsky et al. (1998) propose a clustering of these 220 tags into 42 larger classes and it is this clustered set that was used both in our experiments and those of Stolcke et al. (2000). In measuring the agreement between annotators in labelling this data, Jurafsky et al. (1998) report an average pair-wise kappa of .80 (Carletta et al.,

```
 <Turn Id="utt3" Speaker="A" DA-Type="Open-question"> what do you think was different ten
years ago from now?</Turn>

<Turn Id="utt4" Speaker="B" DA-Type="Statement-opinion"> Well I would say as far as social
changes go I think families were more together.</Turn>

<Turn Id="utt5" Speaker="B" DA-Type="Statement-opinion"> They     did     more     things
together</Turn>

<Turn Id="utt6" Speaker="A" DA-Type="Acknowledge"> Uh-huh</Turn>
```

Figure 2: Excerpt of dialogue from the SWITCHBOARD corpus

1997). An excerpt of dialogue from the SWITCH-BOARD corpus can be seen in Figure 2.

3.2 AMITIÉS

The AMITIÉS project (Hardy et al., 2005) collected 1000 English human-human dialogues from European GE call centres. These calls are of an information seeking or transactional type, in which customers interact with their financial accounts by phone to check balances, make payments and report lost credit cards. The resulting data has been sanitised, to replace identifying features such as names, addresses and account numbers with generic information (*"John Doe"*, *"1 The Street"*) and the corpus is annotated with DAs using XDML, combining slight variant of the 42-class DAMSL (Hardy et al., 2002) with domain specific semantic information such as account numbers and credit card details (Hardy et al., 2003).

The most frequent tag in the AMITIÉS corpus is *Influence-on-listener="Information-request"*, which occurs 20% of the time. For this corpus, the average pair-wise kappa score of .59 was significantly lower than the SWITCHBOARD corpus. For the major categories (questions, answers), average pair-wise kappa scores were around .70. Again, according to the work of Carletta et al. (1997), a minimum kappa score of 0.67 is required to draw tentative conclusions. An excerpt of dialogue from the AMITIÉS corpus can be seen in Figure 3.

4 DA Classification

In this section we briefly describe our approach to DA classification, based solely on intra-utterance features. A key aspect of the approach is the selection of the word n-grams to use as cue phrases. Samuel et al. (1999) investigate a series of different statistical criteria for use in automatically selecting cue phrases, but we use a criterion of *predictivity*,

described below, which is one that Samuel et al. (1999) do not consider.

4.1 Cue Phrase Selection

For our experiments, the word n-grams used as potential cue phrases during classification are computed from the training data. All word n-grams of length 1–4 within the data are considered as candidates. The phrases chosen as cue phrases are selected principally using a criterion of *predictivity*, which is the extent to which the presence of a certain n-gram in an utterance is predictive of it having a certain dialogue act category. For an n-gram n and dialogue act d, this corresponds to the conditional probability: $P(d \mid n)$, a value that can be straightforwardly computed. For each n-gram, we are interested in its *maximal* predictivity, i.e. the highest predictivity value found for it with any DA category. This set of n-grams is then reduced by applying thresholds of predictivity and occurrence, i.e. eliminating any n-gram whose maximal predictivity is below some minimum requirement, or whose maximal number of occurrences in any category falls below some threshold value. This thresholding removes some low frequency, high predictivity n-grams that skew classification performance. The n-grams that remain are identified as our cue phrases. The threshold values that are used in all experiments were arrived at empirically, using a validation set to automatically set the threshold levels independently of the test data, as described in Webb et al. (2005b).

4.2 Using Cue Phrases in Classification

To classify an utterance, we identify all the word n-grams it contains, and determine which of these has the highest predictivity of some dialogue act category (i.e. is performing as some cue). If multiple cue phrases share the same maximal predictivity, but predict different categories, we select the

```
 <Turn Id="2.1" Speaker="Operator" Info-level="Communication-mgt"
Conventional="Opening">good morning customer services sam speaking</Turn>

<Turn Id="3.1" Speaker="Customer" Info-level="Communication-mgt"
Conventional="Opening">erm good morning</Turn>

<Turn Id="3.2" Speaker="Customer" Info-level="Task"
Forward-function="Explanation">erm I was away for about two months and i came back
and my card i don't know whether i have lost it or it is stolen</Turn>

<Turn Id="4.1" Speaker="Operator" Understanding="Backchannel"
Response-to="T3.2">right okay</Turn>

<Turn Id="4.2" Speaker="Operator" Info-level="Task"
Influence-on-listener="Info-request-explicit">can you confirm your name
for me please</Turn>
```

Figure 3: Excerpt of dialogue from the AMITIÉS GE corpus

DA for the phrase with the highest frequency. If the combination of predictivity and occurrence count is insufficient to determine a single DA, then a random choice is made amongst the remaining candidate DAs. If no cue phrases are present, then a default tag is assigned, corresponding to the most frequent tag within the training corpus.

Our best reported figures on the 202k utterance SWITCHBOARD corpus are a cross-validated score of 69.09%, with a single high score of 71.29%, which compares very favourably with the (not cross-validated) 71% reported in Stolcke et al. (2000) for the same corpus. We also presented information that shows that adding a sequence model of DA progressions - an n-gram model of DAs - results in no significant increase in performance (Webb et al., 2005a). This is surprising considering that Stolcke et al. (2000) report their best figures when *combining* a HMM model of the words inside utterances with a tri-gram model of the Dialogue Act sequence, as in the work of Reithinger and Klesen (1997). When Stolcke et al. (2000) add the sequence model to the HMM language model, it adds around 20% points to the final accuracy score over the SWITCHBOARD data.

However, our observation is confirmed by both Serafin and Eugenio (2004) and Ries (March 1999). On the basis of this result, we hypothesise that our cues are highly predictive of dialogue structure, and that much dialogue processing may take place at a very shallow level.

5 Cross-Domain Classification

The central purpose of this paper is to examine the use of automatically extracted cues to tag data *other* than the corpus from which they are derived. The hypothesis we wish to test is that these cues are sufficiently general to work as a classification device on a corpus from a different domain, even containing interactions of a different conversational style. Specifically, SWITCHBOARD is an open domain spoken human-human conversational corpus and we have shown state-of-the-art tagging performance over this data using our cue-based model. We now wish to see how well these same cues perform over the AMITIÉS GE corpus of spoken *t*ask-based dialogues. The dialogues in the AMITIÉS GE corpus are far more goal directed, and contain domain specific cues not found in the general conversational SWITCHBOARD corpus.

The ability to apply cues extracted from one corpus to new data is an interesting challenge. It could confirm work which indicates the prominence of such word cues in language (Hirschberg and Litman, 1993). A tag mechanism that can operate across domains presents a range of benefits - for example it can be used to annotate or partially annotate new data collections.

5.1 DA Mapping

Cross-corpus classification would be simplified if both corpora were annotated with identical DA taxonomies. In actuality, the SWITCHBOARD corpus and the AMITIÉS GE corpus are annotated with *variants* of the DAMSL DA annotation scheme. In the SWITCHBOARD corpus, the hierarchical nature of the DAMSL schema has been flattened and clustered, to produce 42 major classes. In the AMITIÉS GE corpus, the dialogue level schema has been left largely untouched from the DAMSL original. In or-

der to be able to compare automatic classification performance across the two corpora, a mapping is required between the 42-class schema of SWITCH-BOARD and the DAMSL-like XDML schema of the AMITIÉS GE corpus. In their work, Jurafsky et al. (1998) include such a mapping between SWITCHBOARD and DAMSL that covers approximately 80% of the labels in the SWITCHBOARD corpus. We have adapted this slightly to cover minor differences between the XDML used in the AMITIÉS GE corpus and the original DAMSL, although this leaves us with two issues that we need to address.

First there are differences in granularity on both sides. Importantly, in many instances we may identify the most salient role of the utterance, but miss modification information which may make little interpretative difference. For example, mark-up in the AMITIÉS GE corpus makes the distinction between <Forward-function="Assert"> and <Forward-function="Reassert">, whereas mark-up in the SWITCHBOARD corpus ignores such a distinction, and annotates both as type <Forward-function="Assert"> - although the SWITCH-BOARD corpus captures the difference between assertions that are opinions, and those that are not, whereas the original DAMSL does not capture this distinction. To address this mismatch we create a set of super classes by relating the annotations of SWITCHBOARD-DAMSL and the AMITIÉS GE-XDML corpora at the most salient level, according to the mapping contained in Jurafsky et al. (1998). Whilst the majority of tags have a one-to-one correlation, there are elements of both the Forward-Looking Function (see Figure 4) and Backward-Looking Function (Figure 5) that require mapping in both directions.

Secondly, there are a number of AMITIÉS GE tags that we know a-priori we have little or no chance to recognise. For example, the AMITIÉS GE corpus is meticulously annotated to include that certain utterances are perceived as answers to prior utterances. Our approach to DA tagging is purely *intra*-utterance, taking no account of the wider discourse structure, so will not recognise these distinctions. Although such a model of discourse structure should be trivial, based for example on an adjacency pair approach, this will be evaluated further in future work.

5.2 Evaluation Criteria

These issues require that we create two evaluation criteria for our subsequent experiments - **strict** and **lenient**. With strict evaluation, we are required to match *all* elements of the AMITIÉS GE corpus annotation - despite knowing in advance that this is not possible for a range of utterances. We use our strict evaluation criteria to establish a lower bound of performance for our classifier. Our lenient approach is a back-off model, where we require that we correctly identify the most critical part of the multi-part annotation - those that are identified as the most salient.

We'll use the dialogue excerpt shown in Figure 3 as an example of how these two scoring mechanisms work. The first utterance (2.1) is marked as <Info-level="Communication-mgt" Conventional="Opening">. This has a one-to-one correlation with the SWITCHBOARD-DAMSL tag <conventional-opening>. In the case of this example, and in all instances in the AMITIÉS GE corpus, utterances are marked as <Info-level="Task">, *unless* they are from a small set of exceptions, including openings, closings or backchannels, that are annotated as <Info-level="Communication-mgt">. Once an utterance is tagged as one of these exceptions, we know to change the <Info-level> assignment accordingly.

There will be no difference between our strict and lenient evaluation models for the interpretation of this utterance. The same is true for the second (3.1) utterance annotation, which has a direct correlation with SWITCHBOARD-DAMSL annotations. However, the fourth utterance (4.1) includes a <Response-to="T3.2"> annotation that we will not be able to identify using our intra-utterance model. This utterance will be judged correct using the lenient model, and incorrect using the strict metric.

The third utterance (3.2) is marked as <Forward-function="Explanation">. Using the Forward-function map shown in Figure 4, we see that this maps to the super class <Forward-function="Assert">, that in turn maps to the SWITCHBOARD-DAMSL tags <statement-non-opinion> and <statement-opinion>. This means that any utterance identified by the presence of a cue phrase as either <statement-non-opinion> or <statement-opinion> will in fact be tagged as <Info-level="Task" Forward-function="Assert">. Whilst this annotation

$$\left.\begin{array}{l} Forward - function = \text{``}Assert\text{''} \\ Forward - function = \text{``}Reassert\text{''} \\ Forward - function = \text{``}Explanation\text{''} \\ Forward - function = \text{``}Rexplanation\text{''} \\ Forward - function = \text{``}Expression\text{''} \end{array}\right\} Forward - function = \text{``}Assert\text{''} \left\{\begin{array}{l} statement - non - opinion \\ statement - opinion \end{array}\right.$$

Figure 4: Partial Forward-Looking Function mapping table (XDML } SUPERCLASS { SWITCHBOARD-DAMSL)

$$\left.\begin{array}{l} Inf - on - list = \text{``}Info - req - explicit\text{''} \\ Inf - on - list = \text{``}Info - req - implicit\text{''} \\ Inf - on - list = \text{``}Conf - req - implicit\text{''} \\ Inf - on - list = \text{``}Conf - req - explicit\text{''} \end{array}\right\} \begin{array}{c} Influence - on - listener = \\ \text{``}Information - request\text{''} \end{array} \left\{\begin{array}{l} yes - no - question \\ wh - questions \\ open - questions \\ or - clause \\ declarative - question \\ tag - question \end{array}\right.$$

Figure 5: Partial Backward-Looking Function mapping table (XDML } SUPERCLASS { SWITCHBOARD-DAMSL)

captures the salient behaviour of the utterance, it is not an exact match to the original AMITIÉS GE corpus annotation and correspondingly when scoring the lenient model will score this as correct, whereas the exact model will not.

The same is true with the fifth utterance (4.2), annotated in this case as <Influence-on-listener="Info-request-explicit">. A classifier trained over the SWITCHBOARD corpus would identify this (through the mapping see in Figure 5) as <Influence-on-listener="Information-request">, which would be scored as correct using the lenient measure, and incorrect using the exact.

5.3 Classification Experiments

The results of our experiments are summarised in Figure 6. First, to establish our baseline tagging performance, we take the classification algorithm outlined earlier in Section 4, and apply it to the SWITCHBOARD corpus for both training and testing, replicating the work reported in Webb et al. (2005a). In this case, 198,000 utterances are used for training, and a separate 4,000 utterances are used for testing. We achieve a cross-validated score of 69.6%, where the most frequent tag in SWITCHBOARD, <statement-non-opinion>, occurs 36% of the time. This is a confirmation of the work reported in Webb et al. (2005a), and demonstrates that this simple model works exceptionally well for this task.

For the first of the new experiments to test our hypothesis, we substitute the AMITIÉS GE corpus

for the SWITCHBOARD corpus in both steps - training and testing - which will give us an upper bound of performance of this particular classification algorithm over this data. In this experiment, we used 10% of the corpus for testing - giving us a total of 27,000 utterances for training and 3,000 utterances for testing. For all experiments where AMITIÉS GE data is used as a test corpus, both strict and lenient scoring will be used. Strict scoring sets a lower bound for this exercise, and should be greater than chance, which corresponds to the distribution of the most frequent DA tag in each corpus. For strict scoring, where we are required to match all the elements of the AMITIÉS GE XDML tag, we score 65.9% accuracy in this experiment. For lenient, where we must match only the most salient features, we score 70.8% accuracy. Whilst there is no direct comparison to other work on this corpus, Hardy et al. (2005) show partial results for DA classification on this task, looking only at a few major classes, and achieve a score of 86%. However, this includes only the 5 most frequent DA categories, and considers utterances shorter than a certain number of words.

Finally, we attempt cross-domain classification: First, we train our classifier using SWITCHBOARD data, and test using AMITIÉS GE data. We recorded a strict evaluation score of 39.8% tagging accuracy. Using the lenient score, we achieve around 55.7% accuracy. This can be considered a very good result, given the lower bound score of 20% - that is the count of the most frequent tag.

Training corpus	Training utterances	Testing corpus	Test utterances	Common tag (%)	Lenient score	Strict score
SWITCHBOARD	198,000	SWITCHBOARD	4,000	36%	n/a	69%
AMITIÉS GE	27,000	AMITIÉS GE	3,000	20%	70.8%	65.9%
SWITCHBOARD	198,000	AMITIÉS GE	30,000	20%	55.7%	39.8%
AMITIÉS GE	27,000	SWITCHBOARD	198,000	36%	48.3%	40%
SWITCHBOARD	27,000	AMITIÉS GE	3,000	20%	53.2%	38%

Figure 6: Experimental Results

Then we apply the classification in reverse - we train on AMITIÉS GE data, and test on the SWITCHBOARD corpus, using all available data in both cases. Using the strict evaluation metric, we achieve a score of 40.0%, and a lenient score of 48.3%. This compares to a baseline of 36%, so is not a drastic improvement over our lower bound. Some inspection of the data informed us that the AMITIÉS GE data did not include many <backchannel> utterances, so subsequently most of these instances in the SWITCHBOARD corpus were missed by our classifier. By changing the default tag to be <backchannel>, rather than the most frequent tag for the training corpus, we achieve a performance gain to 47.7% with strict scoring, and 56.0% with the lenient metric.

For the last experiment, we also wanted to study the effect of limiting the training data on cross-domain classification, by reducing the SWITCHBOARD data to match that of the AMITIÉS GE training set - that is, to use only 27,000 utterances of the SWITCHBOARD corpus as training data to extract cues, which are then applied both to itself (for reference), and to the AMITIÉS GE corpus. On a related note, part of the work conducted in Webb et al. (2005a) studied the impact of different size training models when classifying SWITCHBOARD data, using models of 4k, 50k and 202k utterances. Whilst substantial improvement was seen when moving from 4k utterances to 50k utterances, the subsequent increase from 50k to 202k utterances had a negligible impact on classification accuracy. With the reduced SWITCHBOARD training set, we score 53.2% with the lenient metric, and 38% with strict, indicating that the reduction is size of the training data has some effect on classification accuracy.

6 Discussion, Future Work

We have shown that the cues extracted from the SWITCHBOARD corpus can be used to success-fully classify utterances in the AMITIÉS GE corpus. We achieve almost 80% of the upper baseline performance over the AMITIÉS GE corpus, when judged using our lenient scoring mechanism - scoring 55.7% using the cross-domain cues, compared to the 70.8% when using in-domain cues. When using the strict measure we still achieve around 60% of the upper bound performance, both results being a substantial improvement over the baseline measure of 20%, corresponding to the most frequent tag in the AMITIÉS GE corpus. This is a significant result, which confirms the idea that cues can be sufficiently general across domains to be used in classification.

However, whilst the experiment using SWITCHBOARD corpus derived cues to classify AMITIÉS GE data works well, the same is not true in reverse. There are two possible explanations for this result. It could be related to the size of data available for training, although our experiments in this area seem to suggest otherwise and so we believe that the composition of the training data is a more crucial element. Although the DA distribution in the SWITCHBOARD corpus is uneven, there is sufficient data for the major classes to be effective on new data that also contains these classes. Although the AMITIÉS GE contains a lot of questions and statements, there is very little of the other significant categories, such as <backchannels>, a key DA in the SWITCHBOARD corpus and conversational speech in general. Correspondingly, the cues derived from the AMITIÉS GE data perform well on a selection of utterances in the SWITCHBOARD corpus, but very poorly on others. We want to perform an in-depth error analysis to see if the errors we obtain in classification accuracy are consistent. We can also compare our list of automatically derived cues phrases, particularly those that overlap between the two corpora, to those reported in prior literature. It might be interesting to see if more complex models, derived using state-of-the art ma-

chine learning approaches, could demonstrate similar portability - i.e is it the simplicity of our model that allows for the observed robust portability?

Finally, we wish to combine SWITCHBOARD and AMITIÉS corpora in the cue learning phase, to see how this effects classification, and apply the results to a range of other corpora, including the ICSI-MRDA corpus (Shriberg et al., 2004).

References

Bunt, Harry. 1994. Context and dialogue control. *THINK*, 3:19–31.

Carletta, J. C., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23:13–31.

Cohen, P. R. and C. R. Perrault. 1979. Elements of a plan based theory of speech acts. *Cognitive Science*, 3.

Core, Mark G. and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, MA.

Grosz, Barbara and Candace Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 19(3).

Hardy, Hilda, Kirk Baker, Laurence Devillers, Lori Lamel, Sophie Rosset, Tomek Strzalkowski, Cristian Ursu, and Nick Webb. 2002. Multi-layered dialogue annotation for automated multilingual customer service. In *Proceedings of the ISLE workshop on Dialogue Tagging for Multimodal Human Computer Interaction, Edinburgh*.

Hardy, H., K. Baker, H. Bonneau-Maynard, L. Devillers, S. Rosset, and T. Strzalkowski. 2003. Semantic and dialogic annotation for automated multilingual customer service. In *Eurospeech, Geneva, Switzerland*.

Hardy, H., A. Biermann, R. Bryce Inouye, A. McKenzie, T. Strzalkowski, C. Ursu, N. Webb, and M. Wu. 2005. The AMITIES System: Data-Driven Techniques for Automated Dialogue. *Speech Communication*, 48:354–373.

Hirschberg, Julia and Diane Litman. 1993. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530.

Jurafsky, Daniel, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1998. Switchboad discourse language modeling project final report. Research Note 30, Center for Language and Speech Processing, Johns Hopkins University, Baltimore.

Prasad, Rashmi and Marilyn Walker. 2002. Training a Dialogue Act Tagger for Humna-Human and Human-Computer Travel Dialogues. In *Proceedings of the 3rd SIGdial workshop on Discourse and Dialogue*, Philadelphia, Pennsylvania.

Reithinger, Norbert and Martin Klesen. 1997. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*.

Ries, Klaus. March, 1999. Hmm and neural network based speech act classification. In *Procceddings of the IEEE Conference on Acoustics, Speech and Signal Processing, volume 1, pages 497–500*, Phoenix, AZ.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1999. Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, Canada*.

Serafin, Riccardo and Barbara Di Eugenio. 2004. FLSA: Extending Latent Semantic Analysis with features for dialogue act classification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Shriberg, E., R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Special Interest Group on Discourse and Dialogue (SIGdial)*, Boston, USA.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Computational Linguistics 26(3), 339–373*.

Tur, Gokhan, Umit Guz, and Dilek Hakkani-Tur. 2006. Model Adaptation for Dialogue Act Tagging. In *IEEE Spoken Language Technology Workshop*.

Webb, Nick, Mark Hepple, and Yorick Wilks. 2005a. Dialogue Act Classification Based on Intra-Utterance Features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*.

Webb, Nick, Mark Hepple, and Yorick Wilks. 2005b. Empirical determination of thresholds for optimal dialogue act classification. In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*.